

FINITE-SAMPLE JUSTIFICATIONS OF MLE:

On the importance of the score function in
estimation and testing

Michael Schweinberger

Department of Statistics

Outline

- asymptotic properties of MLE
- asymptotic versus finite-sample properties
- finite-sample justification of MLE: sufficient statistics
- finite-sample justification of MLE: Godambe (1960)
- relation to statistical testing
- other issues.

Notation

- sample space \mathcal{X}
- sample point $x \in \mathcal{X}$
- family of probability measures P_θ
- admitting a probability density $p_\theta = dP_\theta/d\mu$ (Radon-Nikodym derivative) with respect to a σ -finite measure μ which is indexed by some parameter $\theta \in \Theta \subset \mathfrak{R}$.

Asymptotic properties of MLE

Under regularity conditions, the MLE $\hat{\theta}_n = \hat{\theta}_n(X)$ of θ is

- Wald-consistent: $\hat{\theta}_n \xrightarrow{\text{P}} \theta$ as $n \rightarrow \infty$.
- Fisher-efficient: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{L}} N(0, I(\theta)^{-1})$.

Asymptotic $n \rightarrow \infty$ properties versus finite-sample properties

In practice, which is all that matters:

- n is finite and fixed:

"In practice, the sample size doesn't go anywhere" (Geyer, 2004).

- asymptotics: no statement about any particular element of the sequence $\hat{\theta}_1, \hat{\theta}_2, \dots$.
- what is "large".
- main reason for popularity of asymptotics: finite-sample properties not well-understood.

...and last but not least:

"R.A. Fisher: The Founder of Modern Statistics" (C.R. Rao, 1992).

Finite-sample properties of MLE and sufficient statistics

Log-likelihood ratio $l(\theta_1, \theta_2 | x) = \log L(\theta_1 | x) - \log L(\theta_2 | x)$, statistic $T = T(x)$ ("data summary").

Definition (Rao, 1965, 1994): T is sufficient for θ iff $l(\theta_1, \theta_2 | T) = l(\theta_1, \theta_2 | x)$.

Fisher (1922), Neyman (1935): T is sufficient for θ iff there exist non-negative functions g_θ, h such that

$$p_\theta(x) = g_\theta(T) h(x)$$

thus

$$L(\theta | x) \propto g_\theta(T)$$

and the maximum likelihood estimator solves

$$\frac{\partial \log L(\theta | x)}{\partial \theta} = 0 \quad \text{thus} \quad \frac{\partial \log g_\theta(T)}{\partial \theta} = 0$$

and is hence some function of T , holding for all n .

Example: exponential family:

$$h(x) \exp [\phi(\theta)T] / c(\theta)$$

Likelihood is proportional to

$$L(\theta | x) \propto \exp [\phi(\theta)T] / c(\theta)$$

$$\log L(\theta | x) \propto \phi(\theta)T - \log c(\theta).$$

Let $\phi(\theta) = \theta$, then the maximum likelihood estimator solves

$$\frac{\partial}{\partial \theta} [\theta T - \log c(\theta)] = 0$$

$$\frac{\partial}{\partial \theta} \log c(\theta) = T.$$

Example exponential family: X_1, \dots, X_n i.i.d. Poisson(θ).

$$p_{\theta}(x) = \underbrace{e^{-n\theta} \theta^{\sum x_i}}_{g_{\theta}(T = t)} \underbrace{\left[\prod x_i! \right]^{-1}}_{h(x)}$$

thus $T = \sum X_i$ is sufficient for θ .

Maximum likelihood estimate $\hat{\theta}$ of θ :

$$L(\theta | x) \propto e^{-n\theta} \theta^{\sum x_i}$$

and

$$\frac{\partial \log L(\theta | x)}{\partial \theta} = 0 \Rightarrow \hat{\theta} = \frac{t}{n}.$$

Regular estimating functions

Paper to be discussed: *Godambe (1960) "An optimum property of regular maximum likelihood estimation"*. *Annals of Mathematical Statistics*.

Estimating function $g_{\theta}(x)$ that is "smooth" ("Cramér function", see Neyman, 1959, Cramér, 1946, p. 500).

Estimate $\hat{\theta} = \hat{\theta}(x)$ of θ solves $g_{\theta}(x) = 0$.

Implications:

- no assumption about $p_{\theta}(x)$ other than that it exists and is "smooth".
- numerical implementation: root-finding problem.

Examples

In chronological order:

- Gauss-Markov (Legendre, 1805, Gauss, 1809):

$$g_{\theta}(x) = \frac{\partial}{\partial \theta} \int [x_0 - f_{\theta}(x_1, x_2, \dots)]^2 p_{\theta}(x) d\mu(x) = 0.$$

- method of moments (Pearson, 1894):

$$g_{\theta}(x) = \int x^k p_{\theta}(x) d\mu(x) - x^k = 0.$$

- maximum likelihood (Edgeworth, 1908, 1909, Fisher, 1912):

$$g_{\theta}(x) = \frac{\partial}{\partial \theta} \log p_{\theta}(x) = 0.$$

Fisher (1912, p. 157): "*The most probable set of values for the θ 's will make P a maximum.*"

$P = \log$ -likelihood function.

- minimum chi-square (Smith, 1916):

$$g_{\theta}(x) = \frac{\partial}{\partial \theta} \chi^2(x, \theta) = 0.$$

- general forms of the mentioned methods and most other "reasonable" frequentist methods producing regular estimators.
- not restricted to "continuous" or "discrete" case.

Theorem

A. For all "smooth" g_θ ,

$$\frac{E[\{g_\theta(x)\}^2]}{\left\{E\left[\frac{\partial g_\theta(x)}{\partial \theta}\right]\right\}^2} \geq \frac{1}{E\left[\left\{\frac{\partial \log p_\theta(x)}{\partial \theta}\right\}^2\right]}$$

B. The lower bound is obtained by the choice $g_\theta^*(x) = \frac{\partial \log p_\theta(x)}{\partial \theta}$.

C. Thus, for all "smooth" g_θ ,

$$\frac{E[\{g_\theta^*(x)\}^2]}{\left\{E\left[\frac{\partial g_\theta^*(x)}{\partial \theta}\right]\right\}^2} \leq \frac{E[\{g_\theta(x)\}^2]}{\left\{E\left[\frac{\partial g_\theta(x)}{\partial \theta}\right]\right\}^2}$$

Proof (A)

First moment of Fisher score function:

$$\begin{aligned} & E \left[\frac{\partial \log p_{\theta}(x)}{\partial \theta} \right] \\ &= \int_{\mathcal{X}} \frac{\partial \log p_{\theta}(x)}{\partial \theta} p_{\theta}(x) d\mu(x) \\ &= \int_{\mathcal{X}} \frac{1}{p_{\theta}(x)} \frac{\partial p_{\theta}(x)}{\partial \theta} p_{\theta}(x) d\mu(x) \text{ ("chain rule")} \\ &= \int_{\mathcal{X}} \frac{\partial p_{\theta}(x)}{\partial \theta} d\mu(x) \\ &= 0 \end{aligned}$$

follows from differentiating both sides with respect to θ :

$$\int_{\mathcal{X}} p_{\theta}(x) d\mu(x) = 1.$$

Proof (A) (continued)

Cramér function: without loss of generality,

$$E[g_\theta(x)] = \int_{\mathcal{X}} g_\theta(x) p_\theta(x) d\mu(x) = 0.$$

Differentiating both sides with respect to θ and interchanging the order of differentiation and integration gives

$$\int_{\mathcal{X}} \frac{\partial}{\partial \theta} [g_\theta(x) p_\theta(x)] d\mu(x) = 0$$

$$\int_{\mathcal{X}} \frac{\partial g_\theta(x)}{\partial \theta} p_\theta(x) + g_\theta(x) \frac{\partial p_\theta(x)}{\partial \theta} d\mu(x) = 0 \text{ ("product rule")}$$

Proof (A) (continued)

By chain rule: $\frac{\partial \log p_\theta(x)}{\partial \theta} = \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta}$, thus

$$\int_{\mathcal{X}} \frac{\partial g_\theta(x)}{\partial \theta} p_\theta(x) \, d\mu(x) + \int_{\mathcal{X}} g_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) \, d\mu(x) = 0$$

$$\underbrace{\int_{\mathcal{X}} \frac{\partial g_\theta(x)}{\partial \theta} p_\theta(x) \, d\mu(x)} + \underbrace{\int_{\mathcal{X}} g_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) \, d\mu(x)} = 0$$

$$E \left[\frac{\partial g_\theta(x)}{\partial \theta} \right] + \text{Cov} \left[g_\theta(x), \frac{\partial \log p_\theta(x)}{\partial \theta} \right] = 0.$$

Proof (A) (continued)

Cauchy-Schwarz inequality:

$$\text{Var}[g_\theta(x)] \text{Var} \left[\frac{\partial \log p_\theta(x)}{\partial \theta} \right] \geq \left[\text{Cov} \left[g_\theta(x), \frac{\partial \log p_\theta(x)}{\partial \theta} \right] \right]^2$$

$$E[\{g_\theta(x)\}^2] E \left[\left\{ \frac{\partial \log p_\theta(x)}{\partial \theta} \right\}^2 \right] \geq \left\{ -E \left[\frac{\partial g_\theta(x)}{\partial \theta} \right] \right\}^2$$

$$\frac{E[\{g_\theta(x)\}^2]}{\left\{ E \left[\frac{\partial g_\theta(x)}{\partial \theta} \right] \right\}^2} \geq \frac{1}{E \left[\left\{ \frac{\partial \log p_\theta(x)}{\partial \theta} \right\}^2 \right]}.$$

Proof (B), (C): trivial.

Notable facts

- (1) generalizes Frechet-Cramér-Rao inequality (Frechet 1938, Rao, 1945, Cramer, 1946) giving lower bound to the asymptotic variance of unbiased estimators of θ .
- (2) applies to any n , not only "asymptotically as $n \rightarrow \infty$ ".
- (3) mild regularity conditions: "wide range".
- (4) important problem unsolved: asymptotic nature of standard errors: *"unreliable standard errors in small samples undermine the practical value of MLE!"*.
- (5) side remark: related to birth of GEE methods.

Optimality of Fisher score in estimation and testing

Neyman (1959): regular testing function $g_\theta(x)$ (Cramér function):

test $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 \neq 0$

where

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.$$

"What is the optimal testing function satisfying the two minimum requirements":

- (1) impact of plugging-in consistent estimator of nuisance parameter vanishes in the limit as $n \rightarrow \infty$.
- (2) maximum local power in the limit.

Interesting answers:

(1) $g_\theta(x) \perp s_1(\theta)$, where $s_1(\theta)$ is the part of the score vector corresponding to the nuisance parameter.

(2) Let $g_\theta^*(x)$ be a Cramér function such that $g_\theta^*(x) \perp s_1(\theta)$.

Asymptotic local power is maximum iff $g_\theta(x) = s_2(\theta)$, where $s_2(\theta)$ is the part of the score vector corresponding to the parameter of primary interest.

Fact I: if, as consistent estimate, the MLE $\hat{\theta}_1$ is plugged-in for the nuisance parameter θ_1 :

Neyman (1959) test reduces to Rao's (1948) score test.

Fact II: in fact: optimal Neyman (1959) test.

Other literature... and last but not least, R.A. Fisher.

Connections and more

- Pearson (1900) goodness-of-fit test.
- testing moment restrictions.
- Cox (1961, 1962) nonnested test statistics.
- White (1982) information test statistic.
- ...

Economics: Lagrange multipliers as "shadow prices".

Connections to Mahalanobis D^2 , Hotelling T^2 , Fisher's (1936, 1938) linear discriminant analysis.

Importance of score tests

Hacking (1984): Pearson's (1900) goodness-of-fit test among the 20 most important breakthroughs of the twentieth century considering all fields of scientific inquiry.

Rao (2002): "*The dawn of statistical inference*".

Mahalanobis (1933): "...the history of modern statistics may be said to have begun from Karl Pearson's work on the distribution of the χ^2 in 1900".

Fisher (1922): "...of even greater importance is the introduction of an objective criterion of goodness of fit".

Bera and Biliias (2001):

"... one can risk the educated speculation that every optimal test should be based on the score function" (talking not only about large optimality).

Discussion

Estimators optimal in some respect: may not be optimal in other respects:

- Bayesian methods: exact inference.
- efficient versus robust estimators.
- robustness, motivation: (1) outliers, (2) uncertainty about $p_{\theta}(x)$.
- semi-parametric estimators, non-parametric estimators, "non-parametric" Bayes estimators etc.